DOCUMENT RESUME

ED 389 717 TM 024 197

AUTHOR	Scheuneman, Janice; And Others
TITLE	Effects of Prose Complexity on Achievement Test Item Difficulty.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-91-43
PUB DATE	Jul 91
NOTE	59p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Achievement Tests; *Difficulty Level; Goodness of Fit; *Knowledge Level; Multiple Choice Tests; Prediction; *Prose; Psychology; Regression (Statistics); *Test Construction; Test Items; Verbal Tests
IDENTIFIERS	*Graduate Record Examinations; NTE Test of Professional Knowledge

ABSTRACT

To help increase the understanding of sources of difficulty in test items, a study was undertaken to evaluate the effects of various aspects of prose complexity on the difficulty of achievement test items. The items of interest were those that presented a verbal stimulus followed by a question about the stimulus and a standard set of multiple-choice options. Items were selected for study from two tests with differing demands on an examinee's knowledge base, the NTE Communications Skills test (sample size of about 850 examinees) and the Graduate Record Examinations (GRE) Subject Test in Psychology (sample of 1,000). Standard multiple regression analyses and S. E. Embretson's model fitting procedures were used to evaluate the contribution of various complexity factors to the prediction of difficulty. These factors, which included measures of item structure, readability, semantic content, cognitive demand, and knowledge demand, were found to be successful in predicting item difficulty for these items. The immediate usefulness of the results for test development practice, however, are limited by the fact that only a single item type was studied and by the time required to develop the complexity measures. (Contains 1 figure, 14 tables, and 25 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-91-43

ED 389 717

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

EFFECTS OF PROSE COMPLEXITY ON ACHIEVEMENT TEST ITEM DIFFICULTY

Janice Scheuneman
Kalle Gerritz
Susan Embretson

Educational Testing Service
Princeton, New Jersey
July 1991

BEST COPY AVAILABLE

TM024197

Effects of Prose Complexity on Achievement Test Item Difficulty

Janice Scheuneman, Kalle Gerritz

Educational Testing Service

Susan Embretson

University of Kansas

Copyright © 1991. Educational Testing Service. All rights reserved.

Abstract

Better understanding of sources of difficulty in test items would improve the test development process by bringing the functioning of items more under the control of the test developer. To help increase this understanding, a study was undertaken to evaluate the effects of various aspects of prose complexity on the difficulty of achievement test items. The items of interest were those that presented a verbal stimulus followed by a question about the stimulus and a standard set of multiple-choice options. Items were selected for study from two tests with differing demands on an examinee's knowledge base, NTE Communications Skills and GRE Subject Test in Psychology. Standard multiple regression analyses and Embretson's model fitting procedures were used to evaluate the contribution of various complexity factors to the prediction of difficulty. These factors, which included measures of item structure, readability, semantic content, cognitive demand and knowledge demand, were found to be successful in predicting item difficulty for these items. The immediate usefulness of the results for test development practice, however, are limited by the fact that only a single item type was studied and by the time required to develop the complexity measures.

After approximately 75 years of experience with objective group measurement, relatively little is known about the factors that contribute to the observed difficulty of multiple-choice test items. Despite the benefits that would accrue if item difficulty were more readily controlled by the test developer, the preparation of high quality items that function as intended remains largely an art. Indeed, with test content that is largely verbal, even experienced practitioners of this art have been found to be unable to estimate the difficulty of items with any degree of precision, even for a population with which they were familiar (Bejar, 1983). In general, measurement research has provided little guidance in this task of predicting item difficulty.

Psychometrics might be thought of as one of the last bastions of the "black box" approach to psychological theory. On the one hand is the "stimulus," the test item, and on the other is the examinee's response, with little consideration either of the processes used by the examinee in arriving at the response or of the properties of the item and how these properties influence the response processes. In the psychometric tradition, item difficulty has been defined in terms of the performance of a group of examinees or probabilistic models representing their performance. Although both classical and item response (latent trait) theories have proven very useful in addressing a number of measurement problems, in their present state of development, these theories are largely inadequate as sources of explanatory principles for difficulty in test items.

In recent years, however, within the field of cognitive psychology, a number of investigators have studied the processes involved in performing laboratory tasks similar to those used in many aptitude or intelligence tests, such as spatial reasoning tasks and verbal or geometric analogies. Cognitive component analysis involves the identification of component processes in the solution of a particular type of problem (Carroll, 1976; Pellegrino & Glaser, 1979; Sternberg, 1977a, 1977b). Within this field of research are studies that use a method of complexity factors. Complexity factors indicate those characteristics of a task or test item that affect the processes needed for its solution (Embretson, 1983). In these studies, an item is rated on one or more factors that describe the item in terms of underlying theoretical variables. These ratings yield, through the use of a mathematical model, an indicator associated with item difficulty or response time (Bejar & Yocum, 1986; Mulholland, Pellagrino, & Glaser, 1980; Smith & Green, 1985; Stenner, Smith, & Burdick, 1983; Whitely & Schneider, 1981).

Promising as this work is, however, most items in commonly used tests, especially those designed for college students or college-educated adults, differ in important respects from most of the items or item-like tasks that have previously been studied. In particular, the items in these tests have many more sources of complexity and can often be solved by a variety of processes, not all of which are equally difficult to perform. These items require more involvement of metacognitive processes such as (a) determining the nature of the problem to be solved, (b) deciding which performance

components are relevant for solving the item task, (c) selecting a mental representation for information, (d) allocating resources such as time for problem solution, and (e) monitoring solution processes (Sternberg, 1985). Nevertheless, the method of complexity factors may be used with such items, although the connection between process and observed performance must necessarily be less clearly specified. The importance of various item properties can then be inferred from their empirically observed effects on performance rather than from detailed theoretical formulations concerning process (Chalifour & Powers, 1988; Embretson & Wetzel, 1987).

Achievement tests differ from aptitude tests in that they are intended to measure an individual's competency within a domain of knowledge. Nonetheless, cognitive processes, especially metacognitive processes such as those mentioned in the paragraph above, are relevant in responding to achievement test items. Many test items may also differ with regard to both how easy they are to read and comprehend and the nature of the demand placed by the item on the knowledge structure of the examinee.

In this study, we investigated the use of the method of complexity factors with two sets of achievement test items differing in purpose and in their demand on a knowledge domain. Item difficulty was modeled using variables like those identified in previous research by Embretson and Wetzel (1987) with paragraph comprehension items, in addition to other variables that appeared appropriate for the items being examined.

Method

The Items

This study examined only those items which presented a short prose passage followed by a question with five multiple-choice options. The items were selected from those appearing in two forms of the Graduate Record Examination (GRE) Psychology test, which represented a high level of demand on knowledge, and from three forms of the Reading section of the NTE Communication Skills test, which represented a low level of knowledge demand. The GRE Psychology test is designed to assist graduate school committees and fellowship sponsors assess the qualifications of applicants for graduate study in psychology. The NTE Communication Skills test is typically used to determine whether applicants to teacher training programs or for state licensure as teachers possess basic listening, reading, and writing skills. The reading items are administered in a separately timed section of this test.

Items chosen from the GRE Psychology test were those with at least one sentence in addition to the question; items chosen from the NTE test were those with a stimulus passage consisting of no more than one paragraph of expository prose. In both tests, the items were further restricted to those in which the options were presented in standard multiple-choice format. On each of the two GRE forms, 28 items meeting these criteria were identified, for a total of 56 items. Two of the NTE forms yielded 13 items and the third form yielded 12, for a total of 38 items.

Item difficulty values were obtained for samples of approximately 850 examinees for GRE and 1000 examinees for NTE. Examinees in the sample were restricted to those who had reached the last item in the NTE reading section or the last item of interest for this study in the GRE. Difficulty values were thus based only on examinees who had apparently had an opportunity to respond to the items. The range of item difficulty values for the selected items varied from about 20 to 90 percent correct for the NTE and from about 5 to 95 percent correct for GRE. For the regression analyses, a log transformation of the item percent correct values was used, with high values representing easy items. The mean and standard deviation of the transformed values were .77 and .74 respectively for NTE and .18 and 1.25 for GRE, indicating that the GRE items were more difficult on the average, as well as more variable in difficulty.

Complexity Factors

The complexity factors considered in this study related to properties of the text of the passage, stem, and options of the items and to cognitive and knowledge demands made on the examinee. The text properties related to the item structure, semantic content, and readability of the text according to various standard indicators. The sections below describe the variables related to each of the text properties and item demand areas. Preliminary analyses designed to determine the appropriate specification of some of the variables are also presented.

Text Structure. The variables describing the structure of the text included the total number of words, the number of content words, the

number of three-syllable words, the number of syllables, the number of sentences, the number of sentences per 100 words, the number of syllables per 100 words, and the percent of content words. Nouns, verbs, adjectives and adverbs were considered to be content words. In order to determine the number of content words, a list of function words (non-content words) was developed and these words were deleted from the text. The other variables were obtained using the READABLE program (Micro Power & Light, 1984).

Preliminary analyses of the data indicated that for the structure, readability and semantic variables, predictive power was lost by combining information from the passage/stem and options. (For the structure and readability variables, however, passage and stem were not analyzed separately.) For all subsequent analyses, therefore, information for passage/stem and for options was kept separate. The mean and standard deviation of each of the structure variables and their correlation with difficulty are shown in Table 1. The values for options are the sum for the five options in each item.

- -

Place Table 1 about here

- -

In general, the results show that the NTE items studied had longer passages and longer options. The NTE sentences are also longer, as reflected by fewer sentences per 100 words. (For all items in both tests, each option constituted "a sentence," so the number of sentences in the options was always five.) The percent of content words and syllables per 100 words were quite similar for both passage/stem and

options for the two tests. The correlations with difficulty were generally low, with the highest being the percent of content words in the passage/stem for GRE and sentences per 100 words in the passage/stem for NTE.

Readability. A number of readability indices were generated by the READABLE program. In computing these indices, it was sometimes necessary to amend the item options so that each formed a complete sentence. The indices generated by the program were: Coleman, Dale-Chall, Devereaux (ARI), Flesch grade level, Flesch reading ease, Flesch-Kincaid, Fog, and Holmquist. Many of the structural variables above are used in the computation of these indices, but the Dale-Chall and Holmquist indices also reflect the number of words not appearing on the Dale list of 3000 most common words. Except for the Flesch reading ease, all indices are expressed in reading grade levels. In addition, Kucera-Francis word frequencies were obtained for all words and for all content words in the passage/stem and options (Kucera & Francis, 1967). The values used in the analyses for the Kucera-Francis counts were the means of the log frequencies for the passage/stem and options of each item. The means and standard deviations of each of the readability variables and their correlations with difficulty are shown in Table 2. Note that, unlike the other indices, higher values of the Kucera-Francis frequencies and Flesch reading ease indicate easier material.

- -

Place Table 2 about here

- -

In general, the variables showed that the passage/stem of the NTE items was at a higher reading level than the options, although the Kucera-Francis word frequency counts were similar for both. The same was also generally true for the GRE except for those indicators that take word difficulty into account. Specifically, the Holmquist and Dale-Chall grade level indices and Kucera-Francis counts showed that options were at a higher reading level. This reflected the fact that many of the options for GRE Psychology items consisted of only a few words, making them structurally simple, but those words were often quite technical, such as "retroactive inhibition" or "systematic desensitization."

The correlations with difficulty were generally higher for the readability than for the structural variables; the highest were the Dale-Chall index for the passage/stem for GRE and the Fog index for the passage/stem for NTE. Interestingly, the correlations of the indices for the total item (not shown) were consistently in the expected direction; that is, more difficult readability levels were also associated with the more difficult items. The correlations between difficulty and readability for passage/stem and between difficulty and readability of options, however, are fairly consistently in opposite directions. Table 2 shows the correlations with item difficulty are mostly in the expected directions for passage/stem readabilities on the GRE and for option readabilities on the NTE.

Semantic Content. Propositional analysis has been suggested as a means of representing the difficulty for processing text by Kintsch and VanDijk (1978). They provided a theory-based approach to the problem of

quantifying the surface structure of text in terms of its meaning. Their approach assumes that the basic units of meaning are propositions, a more psychologically important feature of text than its surface structure.

Propositions are composed of concepts, in which the first element is a predicate or relational concept. Typically a verb will be included or implied in a predicate, but other relations such as negation also constitute a predicate. Predicates relate arguments, which are subjects and objects, and other propositions. During the process of reading, a text will be perceived as coherent if a connection is found between new propositions and those stored in short-term memory. Connectives are those propositions that serve the function of providing connections among other propositions in the text. The final type of propositions are modifiers, the adjectives, adverbs, or phrases that modify arguments or other propositions in the text.

In this study, the text in the test items was parsed and scored using procedures suggested by Bovair and Kieras (1981) and Turner and Greene (1978). Raters were trained in these procedures and the numbers of predicates, modifiers, connectives and arguments were determined separately for the passage, stem, and options of each item. One of the GRE forms and two of the NTE forms were coded by two raters to permit the reliability of the coding process to be estimated.

The first estimates of reliability were found to be unacceptably low, so the coding was reviewed carefully. One important inconsistency was found to be in the labelling of the propositions. For example, the two coders might have parsed a sentence in the same way, but one would

call a particular proposition a connective while the other would call it a predicate. Once these inconsistencies were resolved, the reliabilities were found to be more satisfactory. The items with only one rater were also reviewed and labelling modified to be consistent with the others where necessary.

Reliabilities, the correlations between scores assigned by the two raters, stepped up using the Spearman-Brown formula to the number of items actually contributing to each propositional score, are provided in Table 3. The lowest reliabilities are for the stem, primarily because these were so short that the numbers of propositions in each was always quite small. Where items were scored by two raters, the score assigned was the average of the two scores.

- -

Place Table 3 about here

- -

Preliminary analysis suggested that, as with Embretson and Wetzel (1987), separating the different types of propositions resulted in considerably better prediction than using a total propositional count. In the present study, the effects of leaving the counts for the different parts of the item separate were also investigated. Little predictive power was lost for the GRE items by summing the propositional counts for passage and stem into one variable instead of treating the two variables separately. Further collapsing the variables by summing counts for passage/stem and options, however, resulted in considerable loss in prediction of the difficulty values for both tests. The results for combining passage and stem for the NTE were intermediate; initially

these counts were kept separate. Propositional and argument densities, the number of propositions or arguments divided by the number of words, were also obtained for passage/stem and options. Densities were not obtained separately for passage and stem because the word counts had not been obtained separately for these parts of the items. Means and standard deviations of the propositional variables and their correlations with difficulty are shown in Table 4.

- -

Place Table 4 about here

- -

The results for numbers of propositions again shows that the NTE passages and options tend to be longer than those in the GRE. The results for densities, however, are similar in both tests for passage/stems. For options, connective density is higher for NTE and argument density for GRE. This result reflects the short, structurally simple options for many of the GRE items discussed above with the readability results. In general, the correlations with difficulty tend to be higher for the NTE items. The highest correlations with difficulty for GRE are .16 with number of predicates and -.16 with modifier density. For NTE, a number of variables have correlations with the difficulty larger than .16, the largest a -.28 with argument density.

Cognitive Demand. Because of the differences in the nature of the demands placed by the items in the two tests, different cognitive demand variables were used for each. For the NTE, the taxonomy of reading questions suggested by Anderson (1972) was initially selected. In this

test for adult readers, however, a number of questions were found that did not fit readily into any of his categories. After a careful review of the items, a more extensive taxonomy was developed with five demand level categories reflecting increasing distance from the text presented in the items. These are:

1. Transformation. This category is based on Anderson's first four levels which include paraphrase questions or transformed paraphrase questions, that is, questions where the information requested appears in the text, but may be in different words and presented in a different order. Typical stems for this level are "According to the passage..." and "The passage states..."

2. Induction/Deduction. This category is based on Anderson's two remaining levels. Deduction refers to items where alternative choices are particular instances of a superordinate in the stem. With induction, the particular instances are in the stem and the alternatives consist of superordinate or gist statements. Typical stems are "What is the main point..." and "Which of the following statements best summarizes..."

3. Passage-based Inference. Questions in this category require the examinee to make inferences about material stated in the text. Typical stems are "Which of the following can be inferred...", "The author implies...", and "The passage suggests..."

4. Rhetorical Inference. Questions here concern the intent of the author, the structure of the passage, and other inferences to be made about the passage separate from the information given. Some of the questions classified in this category included "The tone of the passage suggests...", "The author refers to...in order to...", and "Which of the following statements most accurately describes the organization..."

5. Reasoning. These questions ask the examinee to reason from material supplied in the passage. Some of the questions classified in this category include "Which of the following kinds of data would be most useful...", "The author's argument would be most weakened if...", and "Which of the following is most closely analogous to..."

Cognitive demand was represented in the GRE items by two sets of categories. The first was a revision and elaboration of Bloom's taxonomy by Emmerich (1989), which takes into account the more recent findings of cognitive psychology. Not all of the categories were

represented among the items on this test, however, and some could be combined without reducing the prediction of difficulty. The five cognitive processes categories used in the analyses were:

1. Support/Weaken a claim, procedure, or outcome. Substantiate a result (demonstrate, prove, confirm, verify, etc.) or Negate it (cast doubt on, critique, contradict, disprove, etc.)

2. Infer (conclude, induce, deduce, diagnose) or Distinguish (differentiate, contrast)

3. Generalize (plausibly universalize, find common element or ground) or Transfer (analogize, apply, carry over)

4. Problem/solve (calculate, inquire, experiment) or Evaluate (appraise, weigh, compare)

5. Identify a concrete piece of relevant information not given in the stem. Recall (recognize, name, discern, locate, match) or Exemplify (illustrate)

The second set of cognitive demand categories were identified from an examination and classification of the GRE items. Many of these items required a translation of information from one form to another. The passage/stem of these items provided either (a) the depiction of a concrete situation or example or (b) an abstract, general or theoretical statement or position. The options then provided (a) concrete examples, one of which is equivalent in important respects to the concrete passage or is a concrete example of the abstract passage; (b) abstract statements one of which is equivalent to the abstract passage or the abstract underlying principle or explanation of a concrete passage; or (c) a label for the principle, procedure, or entity illustrated in the concrete passage or described in the abstract passage. Thus, a combination of the concrete or abstract passage/stem and the three types

of options--concrete, abstract, or label--created six mode of translation categories. A final category in this set consisted of items that presented some other type of task, making a total of seven variables.

For the cognitive and knowledge demand variables, the coding of all items was done independently by the two senior authors, with disagreements resolved through discussion. The number of items classified into each of the cognitive demand categories and the mean difficulty of the items appearing in each are shown in Table 5. Also in Table 5 is the multiple correlation obtained when the categories were treated as dummy variables in a regression predicting item difficulty.

- -

Place Table 5 about here

- -

The demand level categories for NTE appear to represent largely similar levels of difficulty with the reasoning category having the highest mean or easiest items, somewhat contrary to expectation. The multiple correlation is comparable to the zero-order correlations of many of the structural and readability variables, however. If treated as a continuum, the zero-order correlation of demand level with difficulty is .07.

In contrast, the cognitive demand variables for GRE appear to be substantially related to difficulty. For the process variables, the easiest items appear to be associated with the Generalize/Transfer category and the hardest with Problem-Solve/Evaluate. The mode of translation categories show a progression in the mean difficulty levels

where the concrete passage/stems were easier than the abstract passage/stems with the "other tasks" in between. Both sets of cognitive demand categories have multiple correlations in excess of .30; each accounts for somewhat more than 10 percent of the variance in item difficulty for the GRE items.

Knowledge Demand. The knowledge variables only apply to the GRE, since the NTE items were selected because they made little demand on the examinee's knowledge base. Two sets of categories were again selected to represent the knowledge demand. The first of these was developed through examination of the items and was specific to the psychology content. These knowledge demand levels reflect the level of knowledge of psychology probably required for a correct response to the item. These categories were intended to represent some of the kinds of judgments about item difficulty that test developers can readily make and probably often do make in developing the test. The levels were as follows:

1. Reading Comprehension. All the information required is provided in the item passage, although knowledge of psychology might make the material more comprehensible.

2. Popular. Readers of the popular press concerning psychology, such as in Time magazine or Psychology Today, are likely to know the information required.

3. Basic. Examinees familiar with the material in a basic psychology course should possess sufficient knowledge to answer these questions.

4. Intermediate. Items at this level require more depth of understanding than the Basic level, but do not require knowledge of specific details or facts.

5. Advanced. Items require understanding of more advanced concepts or knowledge of more specific detail than those at the Intermediate level.

The second set of categories concerned the manifest content of the items. Separate bits of information are provided in the passage, stem, and options that together describe the item content. Emmerich (1989) developed a classification system which can serve equally well to describe information in each of the three item parts. The categories, which he refers to as "knowledge aspects," were used according to his recommendations to classify the content of the passage, stem and options of the GRE items. The categories were:

1. Theory. Recognized but not fully accepted belief or organized set of beliefs. Hypothesis, model, paradigm, formulation, approach, etc.

2. Criterion. Qualitative or quantitative standard of acceptability, merit, or of unacceptability. Presence or absence of relevance, plausibility, logicalness, validity, completeness, etc.

3. Procedure. Method, means, usage, format, experimental design, procedural control, method of analysis, etc.

4. Relationships. Relationships can be classified as (a) system relationship--pattern, network, series, hierarchy, syndrome, etc.; or (b) individual relationship--principle, if-then statement, correlation, influence, independence, cause-effect, etc.

5. Entities. Entities may be (a) tangible--objects, events, persons, observations, outcomes, conditions, etc.; or (b) categories--topics, diagnoses, themes, states, types, domains, etc.

6. Language. Definitions, narrative or exposition, discourse.

Table 6 presents the number of items in each of the knowledge level and knowledge aspect categories and the mean difficulty value of the items appearing in each. Not all of the knowledge aspects were represented for each of the three item parts. For example, none of the passages was classified into the criterion category. In other cases, preliminary analyses suggested that categories could be combined with little effect on the multiple correlation. These included the

combination of language and entity categories for passages, and the two relationship categories for stems and options. Table 6 also gives the multiple correlations with difficulty level of the knowledge demand categories, coded as dummy variables.

- -

Place Table 6 about here

- -

The different item parts tended to be concentrated in different knowledge categories. Knowledge aspects for passages were primarily split among theory, procedures, and individual relationships, but nearly half the stems were relationships and nearly 60 percent of the options were entities. This corresponds with the observation above that the GRE stems tended to consist of only a few words at a high level of vocabulary. About half the items were also at the basic knowledge demand level.

For both knowledge aspect and knowledge level, the categories show a good spread in mean difficulty levels. For knowledge aspects, the multiple correlations are similar in magnitude to those of the cognitive demand variables for GRE. Knowledge level, however, is the best predictor of item difficulty of all those considered. The categories are ordered in terms of difficulty much as expected, except that the popular press level items are even easier than the reading only level. Since few items were classified at this level, however, it does not appear to much upset the ordering. If demand level is treated as a continuous variable, the zero-order correlation is .46, alone accounting for 21 percent of the variance in item difficulty. This is the best

single predictor, even as a continuous variable. Because level is more theoretically appropriate as a continuum, it was used in this way in the data analyses. Assigning a value of 5 to the advanced level, 4 to the intermediate, etc., the mean and standard deviation respectively of the knowledge level variable were 3.0 and 1.25.

Summary. The complexity factors can be grouped into three broad categories: text properties, cognitive demands, and knowledge demands (GRE only). The text properties include the item structure variables, the readability indicators and the semantic content factors--numbers and densities of propositions. Cognitive demand is represented by the demand level of the questions posed in the NTE items and by the cognitive process and mode of translation variables for the GRE items. Knowledge demand factors include both knowledge level and knowledge aspect. These are summarized in Figure 1 along with a list of variables for each factor.

- -

Put Figure 1 about here

- -

Data Analyses

After the preliminary analyses used in defining the variables for each of the complexity factors described above, an effort was made to reduce further the number of variables for each of the text property factors (structure, readability, and semantic) to be considered. In all analyses, the separate values for passage/stem and options of a given variable, such as number of words, were treated as a set. This set of two or three separate values is referred to in the discussion as

a compound variable. That is, when the compound variable, number of words, was included in a regression analysis, two separate variables, number of words in the passage/stem and number of words in the options, appeared in the regression equation. Multiple regression analyses were performed including different numbers of compound variables within each of the complexity factors in order to evaluate their relative contributions. Taking into consideration the correlations among variables to avoid colinearity, a "best" set of predictors was selected for each factor separately for the GRE and NTE items.

Models of item difficulty were then specified by sets of one or more of the complexity factors. For example, one model indicated that difficulty could be predicted exclusively from the readability factor. The fit of the different models was then evaluated using two different methods. First, models were evaluated using standard multiple regression procedures. For these analyses, the percent of variance accounted for by the model was the measure of fit. Because the number of items was relatively small for NTE, only models with one or two complexity factors (each with up to eight variables) were considered. For the GRE, interest was focused on the contribution to the prediction of difficulty of the complexity factors beyond that contributed by knowledge level. Models with up to four complexity factors including both knowledge level and knowledge aspect were evaluated for these items.

Second, the models were evaluated using a confirmatory model fitting procedure (Embretson, 1984). This procedure, which is based on the linear logistic latent trait model (Fischer, 1973), evaluates the

fit of models based on different complexity factors and permits testing hypotheses about the contribution of each to the prediction of item difficulty. In this study, only those models identified as promising in the regression analyses were subjected to these procedures. Perhaps because of this pre-selection, all complexity factors evaluated were found to improve significantly the fit of simpler models that did not include that factor. Therefore, indices reflecting effect size were obtained to evaluate the practical importance of the models of interest.

The procedure, implemented with Embretson's LINLOG program, evaluates the fit of a given model with regard to the marginal frequencies of correct responses for each item and numbers of examinees at each score level, yielding a linear-logistic fit statistic. In addition to the complexity models of interest, fit was evaluated for two other models: the null model, which specifies that all items have the same difficulty value, and the Rasch model, which permits all items to differ in difficulty. For the complexity models, the difficulty of each item is specified by one of a set of unique values, where the number of values is equal to the number of variables in the model. For example, for a model that specifies that difficulty is a function of only one complexity factor made up of eight variables, the program will attempt to find the best fit to the item data using only eight difficulty values. Thus, models with more factors permit more distinct difficulty values.

Because it permits as many distinct difficulty values as there are items, the Rasch model will generally provide a better fit to the data than the complexity models based on fewer numbers of variables

representing item properties or cognitive demands. The difference between the linear-logistic values for the Rasch model and the null model (one difficulty value) thus provides a range of fit statistics for the response data provided. This range represents the improvement in fit by using the Rasch model rather than the null model. Similarly, the improvement in fit over the null model can be obtained for the various complexity models. The ratio of the fit improvement of a given model to the fit improvement of the Rasch model yields an effect size statistic analogous to the percent of variance accounted for by the models in the regression analyses. This will be referred to as the percent fit statistic.

Another advantage of the Embretson procedure is that models approaching the number of variables for the Rasch model (one per item) can be meaningfully evaluated. Thus, models consisting of more complexity factors can be evaluated using the model fitting procedures than is the case with the regression methods. (In this study, the number of variables in the regression analyses was not permitted to exceed one half the number of items.) Unfortunately, difficulties were encountered for several of the models for the GRE items, where the program was unable to reach convergence, apparently due to the number of items. The program was therefore run separately for the two GRE forms and linear-logistic fit statistics were combined manually. This procedure probably resulted in somewhat inflated fit statistics, but should not have affected the relative importance of the complexity factors in the various models.

Results

The first step in combining the different models for analysis was to reduce the number of variables. Multiple regression analyses were conducted separately for the GRE and NTE to determine the best combination of variables for each of the complexity factors. Variables selected were those that were not too highly intercorrelated and that together contributed most of the total variance that could be accounted for by that factor. Although the importance of the variables differed, the same sets of variables were found to be the best contributors to the prediction of difficulty for both tests.

The results of these analyses for the text property factors are shown in Table 7. The first column for each test shows the squared multiple correlation for each of the compound variables and their sum for each of the complexity factors. The next column shows the percent additional percent of variance accounted for when the compound variables are entered into a step-wise multiple regression. The third column show the order of entry.

- -

Place Table 7 about here

- -

For the structure factor, the best variable set consisted of two word counts--total words and three-syllable words--and two density measures--sentences per 100 words and percent content words. Together these predicted 21.5 percent of the variance in difficulty on GRE and 13.2 percent on NTE. As with the other factors shown in Table 7, the variance accounted for by the full set of variables was greater than

the sum of the separate contributions of each of the compound variables in the set (for structure, 14.6 percent and 10.5 percent for GRE and NTE respectively).

The order of entry of the variables into the analysis was more consistent for the two tests on the readability factor. The first two variables were the Dale-Chall index and the Fog index followed by the Kucera-Francis frequency counts for content words and then for all words. These two frequency counts were highly correlated, but only the frequencies for content words in the passage/stem were correlated with difficulty (see Table 2), suggesting that the counts for all words may have been acting as suppressor variables. Note that all of these readability indicators include some measure of word difficulty--the Fog index through the number of three-syllable words and the others the frequency of usage. As with the structure factor, the percent of variance in difficulty accounted for by the readability variables was higher for GRE than for NTE, about 30 percent and 17 percent respectively.

For the semantic variables, the separate propositional counts for passage and stem for the NTE items were investigated further. While information was indeed lost by combining passage and stem for arguments and modifiers, the loss was small enough for predicates and connectives to make combining them worthwhile. Finally, when all of the propositional measures were placed into the same regression, both for numbers of propositions and for densities, the connectives were found

to contribute little added variance. In addition, the number of connectives was very highly correlated with the number of arguments. Connectives were therefore eliminated from further analyses.

The propositional variables were found to contribute much more to the prediction of difficulty in the NTE than in the GRE. The number of propositions accounted for 47 percent of the variance in difficulty for NTE (eight variables) but only 8 percent for GRE (six variables). The six density variables accounted for 29 and 6 percent of the variance respectively for NTE and GRE.

The cognitive demand variables and the knowledge level indicator all produced only one value per item, so no further reductions in the number of variables were attempted. The knowledge aspect variable, however, yielded a different categorical value for passages, stems, and options requiring a total of 14 dummy variables to represent them in a regression equation. Analyses suggested that categories could be further collapsed for stems and options when all three item parts were included in the regression. For stems, entity was combined with relationships; for options only the procedures category was kept separate. This reduced the total to nine dummy variables, four each for passages and stems and one for options. Together, these accounted for 25 percent of the variance in item difficulty on the GRE.

Combining Complexity Factors: Regression Analyses

The next step in the analyses was to evaluate the effects of combining pairs of complexity factors. The factors to be combined were the text properties--readability, structure, the number of propositions and the propositional densities--and the cognitive and knowledge

factors, which differed for the two tests. For NTE, the only other factor was the cognitive demand level. For GRE, cognitive process, mode of translation, knowledge level and knowledge aspect were included separately as factors in the analyses.

The results for NTE are presented in Table 8. The best prediction was by those pairs consisting of numbers of propositions and one of the other text property factors, where the percent of variance accounted for ranged from 65 to 68. Level added from 3 to 6 percent to the other variable sets. The readability and structure factor pair and the proposition and densities pair can be seen to be accounting for some of the same variance in item difficulty since the combined effect was somewhat less than their sum, but the others appear to be largely independent.

- -

Place Table 8 about here

- -

For GRE items, interest was not only in the effects of the complexity factors but also in how much those combinations contributed to prediction of difficulty beyond the variance accounted for by knowledge level. Results for one and two factors are shown in Table 9. Above the diagonal are the percents of variance accounted for by the factors and factor pairs alone. The highest percents were generally found in combinations that included readability and knowledge aspect. The highest for any pair was the combination of those two factors, which accounted for 48 percent of the variance in difficulty. Nearly as high, however, was the combination of knowledge aspect and structure, which

together accounted for 47 percent of the variance. Comparing the combined contributions of the factor pairs to those of the separate factors, readability and knowledge appeared to account for some of the same variation in difficulty while structure and knowledge aspect appeared to be fairly independent. Readability and structure also appeared to overlap somewhat, as they did with the NTE.

- -

Place Table 9 about here

- -

Below the diagonal in Table 9 are the percents of variance accounted for by pairs of factors in addition to the knowledge level. Again, the pairs that included readability or knowledge aspect accounted for more of the variation in item difficulty. The highest value, however, was for the pair of knowledge aspect and structure, which accounted for 59 percent of the variance, while the second highest was the pair of knowledge aspect and readability with 55 percent. In general, the gain between the percent accounted for by the pair alone and the pair plus knowledge level tended to be lower for the pairs including the readability factor and higher for those including propositions and densities. In fact, for some of the latter pairs, the increase was more than the 21 percent accounted for by the knowledge demand alone. Consequently, the range of values for percent of variance accounted for was smaller when the pairs were combined with knowledge level, between 36 to 59 percent as opposed to 18 to 48 percent for the pairs alone.

In combining sets of three factors, the highest sets were most often those which included knowledge aspect. Moreover, from a theoretical perspective, the more interesting result is probably the contribution of the various complexity factors to the prediction of difficulty in addition to that contributed by the two knowledge demand measures, knowledge level and knowledge aspect. Consequently, the next set of analyses considered the percent of variance accounted for by pairs of factors in addition to the two knowledge measures. The results are shown in Table 10.

- -

Place Table 10 about here

- -

Here the highest percents of variance accounted for were by those pairs including structure. The highest percent was 66 for the structure and density pair closely followed by 65 for the structure and readability pair. Notice that these values are quite similar to the highest percent of variance accounted for on the NTE. The range of percents was further reduced in this analysis, the lowest now being 54 percent for propositions and cognitive process.

A summary of the percent of variance in item difficulty added by each of the factors is given in Table 11. For most of the factors, the additional percent of variance added beyond the 21 percent accounted for by the knowledge level was less than that accounted for by the factor alone. For mode of translation, propositions, and densities, however, the amount added was more than the amount contributed by the factors alone. When knowledge aspect was also added, the propositions and

densities again added more than they predicted on their own. Here also, the strength of the structure factor becomes apparent when the full 22 percent of variance accounted for by structure alone was added to the combination of knowledge level and knowledge aspect. The shrinkage of percent of variance added beyond the structure factor is also more apparent in this table, although densities added about the same amount beyond both knowledge factors and structure as they predicted alone.

- -

Place Table 11 about here

- -

Model Fitting Analyses

Before evaluating the fit of the various complexity models to the item data, the Embretson model fitting procedures were used to evaluate the contribution of the different variables identified through the regression analyses to make up the text property factors (readability, structure, and semantic content) and the knowledge aspect factor for GRE. Table 12 shows the percent of fit of the model for these factors and the loss of fit that occurred when each of the compound variables (different values for different item parts) was omitted in turn from the model.

- -

Place Table 12 about here

- -

For the GRE Psychology items, the loss resulting from the removal of each compound variable from its factor model was substantial; the smallest loss was approximately 3 percent for argument density. For the

NTE items, however, some of the variables appeared to contribute little to the fit of their respective models. In particular, number of words and modifier density could be removed from their models with essentially no loss of fit. For the structure factor, no one of the variables appeared to be very important in itself for the NTE data. The success of the structure factor in fitting the data appeared to result from the combination of variables, but any three of these appeared to do nearly as well as all four, with the largest loss just under two percent for the percent of content words variable. For propositions and densities, predicates and arguments appeared to account for most of the model fit.

Examining the fit of the models made up of each complexity factor, also shown in Table 12, both propositions and densities resulted in a similar percent of fit for GRE and NTE data, approximately 30 percent for propositions and 20 percent for densities. Modifiers appeared to be more important for the GRE items, however, and arguments less so. Readability and structure appeared to be much more important for GRE items than for than for NTE items. For NTE items, these results are similar to those obtained using regression analysis, shown in Table 7, with propositions appearing somewhat more important in relation to the others. For the GRE data, all models showed better percent fit in these analyses than percent variance accounted for in the regression analyses (perhaps because the models had to be fit separately by form), but the propositions and density factors appeared more important compared to the other complexity factors in the model fitting analyses.

For the NTE items, only five complexity factors were considered. Using the Embretson model fitting procedures, it was possible to

consider the fit of a model made up of all five factors. This model provided 87 percent fit. The importance of each of the factors in providing this fit could then be evaluated by examining the fit of the model if this factor were deleted. The results of this evaluation are given in Table 13. The greatest losses came from the removal of either the readability factor or the propositions factor. That is, a model made up of the four factors other than readability had a 70 percent fit, for a loss of 17 percent. The removal of propositions resulted in a 15 percent loss.

- -

Place Table 13 about here

- -

The next two columns of Table 13 show the loss that occurred when a factor in addition to readability was deleted from the model. If readability was not included in the model, structure provided a more important role in fit to the difficulty data than propositions. Removal of structure resulted in a loss of 24 percent of fit. The next columns show the results for a three factor model deleting propositions and another factor. Here, when propositions were excluded, the most important remaining factor for model fit was densities. A model including only reading, structure, and level accounted for only 34 percent of fit, a loss of 38 percent from a four factor model including Propositions. The final columns show the loss when three factors were deleted including both readability and propositions. Note that level of questioning factor added from 4 to 9 percent for the various combined models as opposed to 2.5 percent fit when it was the only factor

included in the model. Level added more to a model when the readability factor was not included.

For GRE Psychology data, a different procedure was used to evaluate the combined models. For these data, the four factor models that were evaluated had 90 to 97 percent fit so no further factors were added to the models. Here the progression of sucessive models used was the same as that used with the regression analyses. Knowledge level was first added to each of the other factors, then knowledge aspect and then structure. These results are shown in Table 14.

- -

Place Table 14 about here

- -

Here, the most important single factors were readability and structure, but both knowledge aspect and propositions contributed more to fit when knowledge level was included in the models. For three factor models, propositions, readability and structure all added about 20 percent to knowledge level and knowledge aspect. Readability, propositions, and densities all added about 20 percent to the models including knowledge aspect, knowledge level and structure. The best fitting model was one including readability with the other three factors; this model provided 97 percent of the fit provided by the Rasch model. Note that in all these analyses, the role of propositions and, to a lesser extent, propositional densities appeared more important than was the case with the regression analyses.

Discussion and Conclusions

This study demonstrates that complexity factors can successfully predict item difficulty of achievement test items. The two tests studied, NTE Reading and GRE Psychology, were selected because of their differences in the demand placed on the examinees' knowledge. Although results differed in many of the particulars, the same factors were found to be predictive of difficulty (beyond that predicted by the knowledge variables for GRE) and the same variables were found to be predictive within those factors common to the two tests (readability, structure, propositions and densities). These similarities suggest that these factors would also be predictive of the difficulty of similar items in other achievement tests for adult populations.

For the NTE items, where the demand placed by the items on the examinee's knowledge base is low, regression analyses showed that combinations of two factors that included numbers of propositions were adequate to predict about 65 percent of the variation in difficulty. The Embretson model fitting procedures permitted the evaluation of models with more factors, however, and marked improvement in fit was then observed. Together, the five factors--readability, structure, numbers of propositions, propositional densities, and level of questioning--accounted for 87 percent of the fit provided by the Rasch model.

For the GRE Psychology items, models with up to four factors were investigated in both sets of analyses. In the regression analyses, models made up of knowledge level, knowledge aspect, structure, and either readability or density accounted for about 65 percent of the

variance, a result similar to that for NTE with two factors. These same models, however, fit the data more than 90 percent as well as the Rasch model. Models with either propositions or densities had 94 percent fit while the best model, which included readability, had 97 percent fit.

In addition to these basic results, a number of other interesting observations were made. First is the result that for the text property variables, the correlations with difficulty were often in different directions for passage/stems and options. Thus, when a single value was used for each of these variables to represent the item as a whole, the correlations with difficulty were considerably attenuated. This finding may mean that the relationship between these indicators and difficulty has been underestimated in previous studies where this distinction between the parts of the item was not made.

A second observation concerned the readability indices; those indices that were most effective in predicting item difficulty were those incorporating some aspect of vocabulary level. The strength of the readability factor, particularly with the Psychology items, may lie in some complex measure of word difficulty as represented by the combination of the four indices. The structure factor was also surprisingly strong with the GRE items, adding about 20 percent to the prediction of difficulty beyond the two knowledge demand variables in both sets of analyses. For the NTE items, neither factor by itself was as important as with the GRE data. When combined with propositions, however, both readability and structure added substantially to prediction of NTE item difficulty.

The semantic variables, both numbers of propositions and propositional densities, were more clearly important than structure or readability in accounting for difficulty of the NTE items. For the GRE items, the regression analyses suggested that these variables were not too important, whether alone or in combination with other variables. The model fitting analyses, however, suggested that the propositions factor was as important as readability and structure in modelling item difficulty. The reasons for this discrepancy are unclear, but suggest that propositional analyses should not discarded as a useful approach in accounting for difficulty in items requiring substantive knowledge. For the NTE items, which evaluated reading achievement, the propositional analyses appear to be particularly promising. The usefulness of this approach is limited, however, by the time required to provide the coding. Good results have since been obtained in other studies with a simplified coding procedure, which was devised following our experience here (Scheuneman & Gerritz, 1990a,b).

The cognitive demand variables were not among the best predictors for either test. For the NTE items, the cognitive level factor appeared to have more value when other factors were included in the analyses than when it was used alone. In some of the preliminary regression analyses, models were evaluated that included only variables for passages and stems. These analyses suggested that level was much more important in accounting for difficulty when the option variables were absent. Option characteristics may thus in some way capture the nature of the cognitive task for these items. The cognitive process and mode of translation variables were better predictors of difficulty for the GRE

items, but they tended to be subsumed by other predictors when the factors were combined.

The knowledge demand variables for GRE items together contributed 37 percent of the variance in item difficulty. Although knowledge demand was not the primary area of interest in this study, the variables chosen (knowledge level and knowledge aspect) suggest ways of quantifying the knowledge component of items that might lead to better procedures for estimating difficulty in the test development process. Such taxonomies deserve further attention. Other dimensions of the knowledge demand and better definitions or better scale properties for the demand level variable may be fruitful areas for investigation.

Finally, the text properties contributed substantially to the models of item difficulty beyond that contributed by the knowledge demand variables. The best combinations of factors accounted for approximately 30 percent more variance in the regression analyses and 35 to 40 percent better fit in the model fitting analyses. The contributions of factors such as structure, which appeared to be nearly independent of the knowledge variables, suggest that this would not be much reduced even if better variables to represent knowledge were available. If this study were repeated with another type of test, the percent of variance accounted for by text property measures might easily be less than that found here. Given the magnitude of the results, however, it seems likely that, at least for this type of item, the prose measures would again be found to be substantially predictive of item difficulty.

This study has several important implications. First, the results suggest that measures of prose complexity could profitably inform the item writing process if they could be made available in a useful form. Simplifying the coding to this end will be a challenge for future research in this area. Second, if changing text properties can be shown to change item difficulty, new strategies for manipulating item difficulty might be devised. Third, the results suggest that skills or abilities outside the particular area of achievement being tested are demanded by some of the items and hence a lack of such skills rather than a lack of relevant knowledge may sometimes lead to incorrect responses. This implies that prose complexity measures would be useful in studying differential item functioning or group differences in test performance (Scheuneman & Gerritz, 1990a,b). Finally, further hypotheses concerning the processes involved in responding to achievement test items might be elicited from the results of this or similar studies.

References

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. Review of Educational Research, 42, 145-170.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. Applied Psychological Measurement, 7, 303-310.

Bejar, I., & Yocum, P. (1986, June). A generative approach to the development of hidden-figures items (RR-86-20-ONR). Princeton, NJ: Educational Testing Service.

Bovair, S., & Kieras. D. E. (1981). A guide to propositional analysis for research on technical prose (Technical report No. 8). Tucson, AZ: University of Arizona.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ: Erlbaum.

Chalifour, C., & Powers, D. E. (1988, May). Content characteristics of GRE analytical reasoning items (RR 88-7). Princeton, NJ: Educational Testing Service.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Embretson, S. E. (1984). A general latent trait model for response processes. Psychometrika, 49, 175-186.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11, 175-193.

Emmerich, W. (November 1989). Appraising the cognitive features of subject tests (Research Report RR-89-53). Princeton, NJ: Educational Testing Service.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Kucera, & Francis (1967). Computational analysis of present-day American English.

Micro Power & Light Co. (1984). Readability calculations according to nine formulas [Computer program]. Dallas TX: Author.

Mulholland, T., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. Cognitive Psychology, 12, 252-284.

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. In R. J. Sternberg & D. K. Detterman (Eds.), Human intelligence: Perspectives on its theory and measurement. Norwood, NJ: Ablex.

Scheuneman, J. D. & Gerritz, K. (April 1990a). The effect of technical content on gender differences in reading passages. In J. Olson (Chair), Further investigations of gender differences in test performance. Symposium presented at the meeting of the American Educational Research Association, Boston.

Scheuneman, J. D., & Gerritz, K. (1990b). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. Journal of Educational Measurement, 27, 109-131.

Smith, R. M., & Green, K. E. (April 1985). Components of difficulty in paper-folding tests. Paper presented at the meeting of the American Educational Research Association, Chicago.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. Journal of Educational Measurement, 20, 305-316.

Sternberg, R. J. (1977a). Component processes in analogical reasoning. Psychological Review, 31, 356-378.

Sternberg, R. J. (1977b). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of intelligence. New York: Cambridge University Press.

Turner, A., & Greene, E. (April 1977). The construction and use of a propositional text base (Technical Report #63). Boulder, CO: University of Colorado, Institute for the Study of Intellectual Behavior.

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. Applied Psychological Measurement, 5, 383-397.

Table 1

Means, Standard Deviations, and Correlations with Item Difficulty of Structural Variables

		GRE			NTE		
Variable		mean	sd	r diff	mean	sd	r diff
No. Words	p/s	62	30	.08	95	34	.05
	opts	35	21	.01	62	25	.07
No. Content Words	p/s	35	18	.11	52	19	.01
	opts	19	15	-.05	34	16	-.01
No. 3 Syllable Words	p/s	10.1	6.3	-.13	16.6	7.5	.09
	opts	7.1	4.5	.17	11.2	5.6	-.06
No. Sentences	p/s	3.7	1.7	.09	4.6	2.0	-.05
	opts	5.0	0.0	--	5.0	0.0	--
No. Syllables	p/s	100	50	.05	156	58	.06
	opts	59	35	.03	102	41	-.01
% Content Words	p/s	.57	.05	.22	.55	.06	-.09
	opts	.52	.13	-.14	.55	.09	-.18
Sentences per 100 Words	p/s	6.3	1.9	.03	4.9	1.5	-.20
	opts	18.1	7.4	-.02	9.6	4.7	.01
Syllables per 100 Words	p/s	163	16	-.12	163	13	.06
	opts	172	25	.10	166	22	-.09

Table 2

Means, Standard Deviations, and Correlations with Item Difficulty of Readability Indices

		GRE			NTE		
Index		mean	sd	r diff	mean	sd	r diff
ARI	p/s	11.1	3.5	-.12	13.7	4.4	.26
	opts	7.2	3.7	.08	9.1	3.9	-.03
Coleman	p/s	12.2	2.9	-.13	12.5	2.4	.13
	opts	10.2	4.3	.11	11.6	4.3	-.07
Dale-Chall	p/s	11.3	3.1	-.28	11.2	2.2	.01
	opts	12.7	3.6	.04	10.9	3.7	-.11
Flesch Grade Level	p/s	10.0	2.2	-.17	10.6	2.0	.14
	opts	9.4	2.9	.08	9.5	2.7	-.08
Flesch Reading Ease	p/s	51.4	14.4	.13	46.2	13.3	-.21
	opts	54.6	20.2	-.10	54.2	19.0	.05
Flesch Kincaid	p/s	10.4	2.8	-.12	12.4	3.5	.26
	opts	7.4	3.0	.08	8.8	3.1	-.03
Fog	p/s	13.7	3.5	-.21	15.9	3.8	.27
	opts	11.9	3.9	.13	12.4	3.4	-.01
Holmquist	p/s	7.2	1.2	-.22	7.1	0.8	.06
	opts	7.6	1.4	-.06	6.9	1.3	-.11
Kucera-Francis All words	p/s	5.8	0.4	-.03	6.1	0.3	.12
	opts	4.1	1.6	-.02	6.0	0.7	.04
Kucera-Francis Content words	p/s	4.2	0.6	.25	4.6	0.3	.07
	opts	3.1	1.1	.01	4.5	0.7	-.03
Powers	p/s	6.6	0.8	-.13	7.0	0.8	.24
	opts	6.1	1.1	.09	6.3	1.0	-.06

Table 3

Reliability of Propositional Coding

	GRE		NTE	
	28 Items*	56 Items	25 Items*	38 Items
Predicates				
Passage	.98	.99	.92	.95
Stem	.88	.94	.77	.84
Options	.93	.96	.90	.93
Total	.96	.98	.90	.93
Modifiers				
Passage	.97	.98	.94	.96
Stem	.55	.71	.70	.78
Options	.94	.97	.91	.94
Total	.96	.98	.95	.97
Connectives				
Passage	.98	.99	.96	.97
Stem	.70	.82	.74	.81
Options	.91	.95	.92	.95
Total	.96	.98	.94	.96
Arguments				
Passage	.95	.97	.98	.99
Stem	.67	.80	.79	.85
Options	.97	.98	.95	.97
Total	.95	.97	.98	.99

*Reliabilities for passages were based on 19 items for GRE and 16 for NTE because some passages had more than one item. Full length reliabilities were estimated using the Spearman-Brown formula.

Table 4

Means, Standard Deviations, and Correlations with Item Difficulty of Semantic Variables

		GRE			NTE		
		mean	sd	r diff	mean	sd	r diff
No. Predicates	pass*	12.2	6.9	.16	14.0	6.1	.21
	stem				2.8	1.2	-.04
	opts	7.7	4.2	.02	11.6	5.0	.15
No. Modifiers	pass*	8.1	4.6	-.03	10.1	5.7	-.19
	stem				1.5	1.1	.04
	opts	5.0	4.1	-.06	7.1	4.6	-.04
No. Connectives	pass*	7.8	4.9	.06	13.7	6.6	-.05
	stem				1.6	1.3	-.15
	opts	3.2	4.3	.02	8.4	5.0	-.08
No. Arguments	pass*	10.5	4.2	.01	17.7	8.3	-.19
	stem				2.6	1.0	.04
	opts	6.1	4.0	.06	5.5	3.3	-.04
Predicate Density	p/s	.19	.04	.14	.19	.07	.08
	opts	.24	.09	.01	.21	.10	.26
Modifier Density	p/s	.14	.05	-.16	.13	.07	-.23
	opts	.14	.08	-.14	.12	.07	.05
Connective Density	p/s	.12	.04	.08	.16	.06	-.25
	opts	.07	.08	-.09	.14	.07	-.05
Argument Density	p/s	.18	.04	-.04	.22	.07	-.28
	opts	.20	.08	.10	.10	.08	.07

*passage and stem have been combined for GRE items.

Table 5

Number of Items, Mean Difficulty and Multiple Correlations for Cognitive Demand Categories

NTE	N items	mean difficulty	multiple r
Demand Level			.12
Transformation	7	.79	
Induct/Deduct	9	.67	
Inference	8	.75	
Rhetorical	10	.77	
Reasoning	4	.98	
GRE			
Cognitive Process			.33
Support/Weaken	9	.14	
Infer	21	.05	
Generalize	16	.62	
Problem-Solve	3	-1.20	
Identify	7	.19	
Mode of Translation			.37
Concrete/Concrete	1	1.80	
Concrete/Abstract	2	.99	
Concrete/Label	26	.42	
Abstract/Concrete	4	-.13	
Abstract/Abstract	4	-.96	
Abstract/Label	3	-.08	
Other Task	16	.00	

Table 6

Number of Items, Mean Difficulty and Multiple Correlations of Knowledge Demand Categories for GRE

Knowledge Aspects

	Passage		Stem		Options	
	N items	mean diff	N items	mean diff	N items	mean diff
Theory	11	.22	12	.62	1	-.22
Criterion		--	5	-.34	1	1.52
Procedure	18	-.12	9	.69	9	.66
Relationship			25	-.05	10	.14
System	5	-.49				
Individual	17	.53				
Entity	5	.64	3	.28	33	.10
Language[1]			2	-.68	2	-1.04
Multiple r	.29		.32		.29	

Knowledge Level[2]

Advanced	9	-1.24
Intermediate	6	.12
Basic	27	.31
Popular	4	1.45
Reading Comp	10	.62
Multiple r	.56	

[1]Knowledge Aspect Categories, Entity and Language have been combined for Passages.

[2]Knowledge level refers to the total item.

Table 7

Regression Analysis Results for
Structure, Readability and Semantic Variables

	GRE			NTE		
	R^2	% Var. Added	Entered	R^2	% Var. Added	Entered
Structure						
% Cont. words	.073	7.3	1	.033	3.4	2
# 3-syl words	.065	5.9	2	.023	3.8	4
# words	.008	6.9	3	.007	1.8	3
sent/100 words	.000	1.4	4	.042	4.2	1
Total	.146	21.5		.105	13.2	
Readability						
Dale-Chall	.084	8.4	1	.021	2.8	2
Fog	.077	4.3	2	.076	7.6	1
K-F Content	.065	2.9	3	.009	1.2	3
K-F All	.001	14.5	4	.014	5.0	4
Total	.227	30.1		.120	16.6	
Propositions						
Predicates	.025	2.5	1	.055	23.8	2
Arguments	.004	4.9	2	.128	12.8	1
Modifiers	.003	0.7	3	.039	10.4	3
Total	.032	8.1		.222	47.0	
Densities						
Predicates	.019	0.5	3	.069	19.4	2
Arguments	.012	1.4	2	.093	9.3	1
Modifiers	.038	3.8	1	.055	0.6	3
Total	.069	5.7		.217	29.3	

Table 8

Percent of Variance Accounted for
by Pairs of Complexity Factors in NTE

	Props	Dens	Read	Struct	Level
Densities	65				
Readability	67	49			
Structure	68	56	24		
Level	50	32	23	17	
R^2 for one factor	.47	.29	.17	.13	.01

Table 9

Percent of Variance Accounted for by Pairs of Complexity Factors in GRE*

	R	KA	S	MT	CP	P	D
R^2 for one factor	.30	.25	.22	.14	.11	.08	.06
Readability (R)	--	48	40	44	36	34	34
Know. Aspect (KA)	55	--	47	35	34	37	34
Structure (S)	48	59	--	33	29	28	32
Mode Trans. (MT)	49	49	46	--	24	22	19
Cog. Process (CP)	44	43	41	42	--	18	18
Propositions (P)	47	50	42	45	36	--	18
Densities (D)	48	50	46	47	42	42	--
R^2 for one factor and Know. level	.41	.37	.38	.37	.27	.33	.33

*Percents above the diagonal are for pairs of factors; Percent below diagonal are for pairs plus Knowledge Level, which alone accounted for 21 percent of the variance in difficulty.

Table 10

Percent of Variance Accounted for by Pairs of Complexity Factors Plus Knowledge Aspect and Level in GRE

	Struct	Read	Props	Cog Proc	Mode Trans
Density	66	61	58	56	58
Mode Trans.	62	60	60	58	
Cog. Process	61	57	54		
Propositions	61	60			
Readability	65				
% accounted for by factor, KL & KA	59	55	50	43	49

Table 11

Percent of Variance Added by factor Beyond Knowledge Level, Knowledge Aspect and Structure in GRE

		Percent Variance Added by		
	% Var. factor alone	factor +KL	factor +KL +KA	factor +KL +KA+S
Readability	30	20	18	6
Knowledge Aspect (KA)	25	16	--	--
Structure (S)	22	17	22	--
Mode of Trans.	14	16	12	3
Cog. Process	11	6	6	2
Propositions	8	12	13	2
Densities	6	12	13	7
Knowledge Level (KL)	21	--	--	--

Table 12

Model Fitting Analysis Results
Separate Complexity Factors

	GRE		NTE	
	%fit	Loss	%fit	Loss
Structure	38.9		12.0	
% Cont. words		9.4		1.9
# 3-syl words		22.4		1.0
# words		14.7		0.0
sent/100 words		4.3		0.7
Readability	43.7		14.2	
Dale-Chall		4.1		3.4
Fog		9.9		1.2
K-F Content		13.4		4.3
K-F All		17.4		4.4
Propositions	31.5		30.7	
Predicates		14.1		19.3
Arguments		5.3		18.0
Modifiers		6.8		4.4
Densities	20.4		21.3	
Predicates		4.7		13.9
Arguments		2.8		19.6
Modifiers		9.1		0.0
Knowledge Aspect	31.3		--	
Passage		16.5		
Stem		11.6		
Options		6.3		

Table 13

Loss in Percent fit for Models
Excluding Reading and Propositions for NTE

	Model							
	excludes factor		excludes factor, Reading		excludes factor, Props		excludes factor, Rdg., Props	
Factor	% fit	Loss	% fit	Loss	% fit	Loss	% fit	Loss
Reading	70	17	--	--	60	12	--	--
Structure	78	9	46	24	52	20	23	37
Propositions	72	15	60	10	--	--	--	--
Densities	78	9	58	12	34	38	18	42
Level	83	4	61	9	67	5	51	9
All[1]	87		70		72		60	

[1]No "factor" is excluded other than Reading or Propositions as specified.

Table 14

Percent fit for Models Including Knowledge Level, Knowledge Aspect, Structure for GRE

	Model							
	factor alone	factor +KL		factor +KL +KA		factor +KL +KA+S	
Factor	% fit	% fit	% Added	% fit	% Added	% fit	% Added
Readability	44	50	28	74	19	97	22
Knowledge Aspect (KA)	31	55	33	--	--	--	--
Structure(S)	39	53	31	75	20	--	--
Mode of Trans.	23	42	20	64	9	90	15
Cog. Process	15	28	6	66	11	91	16
Propositions	32	56	35	77	22	94	19
Densities	20	47	26	67	12	92	19
Knowledge Level (KL)	22	--	--	--	--	--	--

Figure 1

Summary of Complexity Factors and Variables

	Complexity Factor	Variables
Text Properties	Structure	No. of Words, content words, syllables, 3-syllable words, sentences; Percent content words; Syllables, sentences per 100 words
	Readability	Standard readability indices; Kucera-Francis word frequencies for all words, content words
	Semantic Content	
	Propositions	No. of arguments, connectives, modifiers, predicates
	Density	Density of arguments, connectives, modifiers predicates
Cognitive Demand	Demand level	NTE: Transformation, induction/deduction, passage-based inference, rhetorical inference, reasoning
	Cognitive Process	GRE: Support/weaken, infer, generalize, problem-solve, identify
	Mode of Translation	GRE: Abstract passage with abstract, concrete or label options; Concrete passage with abstract, concrete or label options; Other
Knowledge Demand (GRE only)	Knowledge Level	Reading comprehension, popular, basic, intermediate, advanced
	Knowledge Aspect	Theory, criterion, procedure, relationships, entities, language